A New Corpus Resource for Studies in the Syntactic Characteristics of Terminologies in Contemporary English

Alex C. Fang[1], Jing Cao[2] and Yang Song[2]

Dialogue Systems Group
Department of Chinese, Translation and Linguistics
City University of Hong Kong
Tat Chee Avenue, KowloonHong Kong SAR, PR China
[1]acfang@cityu.edu.hk,
[2]{cjing3,songyang2}@student.cityu.edu.hk

Abstract: In this paper, we present a new corpus resource that has been constructed specially for the study of the syntactic characteristics of terminologies. The corpus is based on the British component of the International Corpus of English (ICE-GB), comprising four parallel subject domains from two text categories (i.e. academic vs. popular prose) with a total of about 200,000 running word tokens. The resource is richly annotated at lexical, grammatical, syntactic, and terminological levels. It is also parameterized according to both text categories and subject domains. The corpus resource is expected to contribute towards a linguistically motivated description of terms and their internal structures. It is also expected to provide an analytical framework for the study of relations between terminological use and text categories as well as subject domains.

Key words: syntactic tree, treebank, syntactic function, terminology, ICE-GB, noun phrase, term annotation, corpus, syntax.

1 Introduction

Automatic term recognition (ATR) and extraction have been a challenging task and encouraged rigorous efforts of researchers from a wide range of backgrounds and disciplines. Nevertheless, past work on terminological extraction tends to focus on specific subject domains, and mainly in the field of biochemistry and medicine such as Ananiadou et al. 2000, Nenadic et al. 2005, Aubin and Hamon 2006, and Ville-Ometz et al. 2007, to name just a few. Some work on other domains such as computing (e.g. Eumeridou et al. 2004; L'Homme 2002; Nakagawa and Mori 2003), economy (e.g. Rodriguez et al. 2007), and legislation (e.g. Ha et al. 2008; Kit and Liu 2008). Those studies are domain specific in a good sense that they concentrate on domain-specific issues like domain knowledge and associated knowledge expressions on the lexical level. Yet they are domain limited in an undesirable sense, which leads to difficulty in evaluating the performance and interoperability of the existing term recognition systems across a set of different domains. Additionally, it remains an issue how such systems will adapt to new domains.

Another noticeable issue is that, among the linguistic features employed in ATR systems, syntactic features have been mainly observed at the phrasal level, and seldom from the perspective of syntactic structures at a clausal level. Grammatical patterns, such as 'noun', 'noun + noun', 'adjective + noun', 'noun + preposition + noun', have been integrated with statistic measurements to determine the termhood (e.g. Frantzi et al. 2000; Pazienza et al. 2005). Eumeridou et al. (2004) go beyond the grammatical patterns and examine how term occurrences correlate the argument structure of verbs across three domains chosen from the British National Corpus. Their findings show an uneven distribution of terms in different argument structures[1], and they also notice the influence that different domains have upon term occurrences. Although the study focuses on the verbal syntax only, it does indicate that syntactic features of terminological entities warrant a worthwhile research topic and that text categories such as registerial types and subject domains should also be a parameter to consider. It is reasonable to believe that further improvement of ATR systems can be achieved by exploring deeper, linguistically motivated analysis of the relation between terminologies and linguistic parameters.

The main focus of this paper is to present a new corpus resource that has been constructed specially for the study of the syntactic characteristics of terminologies. Existing term-annotated corpora are typically domain-specific, such as GENIA (Ohta *et al*. 2002), and typically used as a resource for statistical training. The new corpus resource is different in that it is built on general domains and is richly annotated for syntactic information, especially for detailed annotation of the syntactic categories and their functions within the clause complex that is often dependent on verb sub-categorization. The corpus is based on the British component of the International Corpus of English (ICE-GB), comprising four parallel subject domains from two text categories (i.e. academic vs. popular prose) with a total of about 200,000 running word tokens. The resource is richly annotated at lexical, grammatical, syntactic, and terminological levels. It is also parameterized according to both text categories and subject domains. The tree bank is expected to contribute towards a linguistically motivated description of terms and their associated syntactic structures. It will also provide an analytical framework for the study of relations between terminological use and text types as well as subject domains. The richly annotated trees will facilitate studies in the linguistic relations of terms for the purpose of ontology construction.

In the rest of this paper, we will first of all describe the construction of the corpus, including the selection of the corpus material, the annotation schemes for grammar and syntax, and an inter-annotator analysis of the manual annotation of terms. We shall then report some of our initial empirical observations of the syntactic characteristics of noun phrases (NP) that are terminological entities as opposed to generic NPs across different types and domains. For this purpose, we will describe the distribution of general NPs in terms of text categories and subject domains. We will

[1] In lexical semantic terms, argument structure refers to the semantic type of the verb and its related elements such as agent and theme. The same term is also loosely used in syntax to refer to the sub-categorisation, or valency structure or complémentation type of verbs.

then describe the distribution of terminological NPs according to the same parameters, focusing on their syntactic functions in the tree structure.

2 Corpus Construction

2.1 Corpus resource for term annotation

Our on-going research attempts to extend the previous studies by exploring the syntactic characteristics of terminological entities across different text types and subject domains in contemporary English. To achieve our objectives, the British component of the International Corpus of English (ICE-GB; Greenbaum 1996) was chosen as a basis for the following reasons: First, it is encoded for a variety of text categories and subject domains. Secondly, it is already grammatically tagged, syntactically parsed and manually validated. Finally and most importantly, it is annotated with a rich set of linguistically motivated syntactic relations that will maximally enhance our intended study. The following sections will first describe the resource created from the ICE-GB and introduce its part-of-speech (POS) and syntactic annotations.

2.1.1 Creation of a sub-corpus

The British component of the International Corpus of English (ICE-GB) is a one-million-word corpus comprising both spoken and written British English from the 1990s (Greenbaum 1996; Fang 2007). The spoken section represents 60% of the total size of the corpus with 300 sample texts. The written section accounts for 40% of the corpus with 200 texts. Each component text has about 2,000 word tokens. Table 1 summarizes the text categories in the ICE-GB together with the number of component texts.

Table 1. The structure of ICE-GB

Spoken			**Written**		
Dialogue	Private	100	Non-printed	Student writing	20
	Public	80		Correspondence	30
Monologue	Unscripted	70	Printed	Informational	100
	Mixed	20		Instructional	20
	Scripted	30		Persuasive	10
				Creative	20

Given the purpose of our study, texts from the category of informational writing constitute a suitable source of texts, which is further divided into three sub-categories: academic writing, popular writing and press news reports. Two contrastive text types, i.e., academic writing and popular writing, were chosen. The two text types cover four parallel subject domains comprising ten texts each. Table 2 presents the composition of the sub-corpus created from ICE-GB.

Table 2. The structure of the sub-corpus

Text Type	Subject Domain	Domain Code	# of Texts	# of Words
Academic writing	Humanities	AHUM	10	24,363
	Social sciences	ASOC	10	24,280
	Natural sciences	ANAT	10	24,165
	Technology	ATEC	10	23,386
Popular writing	Humanities	PHUM	10	27,168
	Social sciences	PSOC	10	23,110
	Natural sciences	PNAT	10	23,150
	Technology	PTEC	10	23,584
Total			**80**	**193,206**

As can be seen from Table 2, the sub-corpus comprises 80 texts similar in size with a total number of 193,206 word tokens.

2.1.2 Tree annotations in the ICE-GB

All the texts in ICE-GB are richly annotated grammatically and syntactically (Fang 1996, 2000, 2006, 2007). When the 80 texts from ICE-GB were selected to create the sub-corpus, a treebank was effectively created that comprises 8,306 syntactic trees.

```
PU CL(main,montr,pass,pres)
 SU NP()
  DT DTP()
   DTCE ART(def) {The}
  NPHD N(com,plu) {fibres}
  NPPO PP()
   P PREP(ge) {of}
   PC NP()
    NPHD N(com,sing) {group B}
 VB VP(montr,pres,pass)
  OP AUX(pass,pres) {are}
  MVB V(montr,edp) {found}
 A PP()
  P PREP(ge) {in}
  PC NP()
   DT DTP()
    DTCE ART(def) {the}
   NPPR AJP(attru)
    AJHD ADJ(ge) {autonomic}
   NPPR AJP(attru)
    AJHD ADJ(ge) {nervous}
   NPHD N(com,sing) {system}
 PUNC PUNC(per) {.}
```

Fig. 1 – An example of syntactic annotations in the ICE-GB

As noted in Figure 1 above, the tree structure is richly annotated with fine-grained grammatical and syntactic information. At the grammatical level, words are coded with part-of-speech (POS) tags that include a head tag (such as nouns, verb, and adjectives) with a set of attributes indicating the subcategorizations of the head tag.

For instance, the verb `found` enclosed within a pair of curly brackets is tagged as `V(montr,edp)`, namely, a mono-transitive verb in past participial form. As another example, `{The}` is assigned a label `ART(def)`, meaning it is a definite article, and `{fibres}` is a common noun in its plural form. Syntactically, each node comprises two labels: one representing its syntactic category (such as noun phrase and adjective phrase) and the other the syntactic function. Take the node `SU NP()` as an example, which indicates that it is a noun phrase (`NP`) functioning as the subject (`SU`) of the clause. The same NP comprises a determiner (`DT`), the head (`NPHD`) and a post-modifier (`NPPO`). The definite article `The` constitutes the central determiner (`DTCE`), a daughter node of `DT`. See Appendix for a complete list of all the parsing symbols. With such a system of syntactic categories and their associated syntactic functions, the corpus forms a valuable testbed according to which grammatical relations of various kinds can be investigated. The syntactic framework will also form an informative context within which terms and term relations can be usefully examined.

2.2 Term annotation

Term annotation was carried out manually during a period of four months, and has gone through the following procedures:

- Training of the annotators: The training session helps the annotators get familiar with the special format of the target texts, which are parsed and represented in a form exemplified in Fig. 1.
- Analysis of inter-annotator agreement: This step was taken to establish the consistency and therefore the quality of the annotations by the three different annotators given the same text, and a higher statistic agreement will demonstrate the confidence of the manual annotation.
- Actual annotation: With an annotation guideline, annotators mark up the terms with the help of dictionaries, online dictionaries and term banks.
- Manual examination of terminological annotations.

In the remaining of this section, we shall first describe the annotation guideline and then report the results from the inter-annotator agreement test. The basic statistics of the terminologically annotated corpus resource will be presented in Section 3.

2.2.1 Annotation guideline

Before describing the guideline, we first introduce the operational definition of terminological entities. To our understanding, terms by definition primarily correspond to noun-phrase (NP) groups and thus consist of words that are single nouns or complex noun phrases (Kageura et al. 2004; Nakagawa 2001; Nakagawa and Mori 2003). Following Eumeridou et al. (2004), we also consider terms in a pragmatic sense. Take text `w2a-031` for example. The text is about “blind shaft drilling” under the domain of *technology*. In addition to terms in technology and engineering, we may also mark up terminological entities from related domains such as *environment*. Given such a definition, a working guideline for annotation was made:

- Among the NPs, proper names of places, countries, organizations or institutes are excluded from the current study, and therefore, will not be annotated.
- Variant terms will be annotated.
  - Singular and plural forms of a term will both be regarded as terms in case some termbanks only collect singular form of a term.
  - When an N_1+N_2 compound is a term, the sequence N_2 + of + N_1 will also be treated as a term.
  - Variant spellings of the same term will be accepted.
- With nested terms, we only mark up the longest part as a multi-word term.
- Terms are marked with '<' at the beginning and '>' at the end in the tree diagram, and the resulting NP is described by an additional attribute '`term`'. See Figure 2.

```
PU CL(main,montr,pass,pres)    A PP()
 SU NP(term)                    P PREP(ge) {in}
  DT DTP()                       PC NP(term)
   DTCE ART(def) {The}            DT DTP()
  NPHD N(com,plu) {<fibres>}       DTCE ART(def) {the}
  NPPO PP()                       NPPR AJP(attru)
   P PREP(ge) {of}                 AJHD ADJ(ge) {<autonomic}
   PC NP()                        NPPR AJP(attru)
 NPHD N(com,sing) {group B}        AJHD ADJ(ge) {nervous}
                                  NPHD N(com,sing) {system>}
```

Fig. 2 – Examples of term annotations in the tree structure

2.2.2 Inter-annotator agreement

Three annotators were trained to mark up terms. All the three annotators are university students majoring in linguistics. Among them, two are undergraduates who have been admitted to postgraduate study and one is a PhD candidate. To measure the inter-annotator agreement, two texts were taken from the pre-selected sub-corpus from ICE-GB, with a total number of about 4,000 words. During the annotation stage, the annotators were allowed to refer to the guideline or other sources such as online termbanks and dictionaries, in addition to their linguistic knowledge. They were not allowed to confer with each other over the annotation.

We then compared the annotations among the three annotators by using *F* score, which is considered to be a standard measure to determine the inter-annotator agreement (Corbett et al. 2007) and has been commonly used in previous studies (see, for example, Demetriou and Gaizauskas 2003, Morgan et al. 2004, Vlachos and Gasperin 2006 and Kolarik 2008). Therefore, the inter-annotator agreement was computed pair-wise using a measure defined in (1):

$$F = \frac{2 \times C \times 100}{M_1 + M_2} \tag{1}$$

where M_1 and M_2 are the number of markable terms in a given text marked up by Annotators 1 and 2 respectively, and *C* is the total number of times both annotators agree on a markable term in that same text. To calculate the *F* score, the total number of terms marked by annotators A, B, and C were counted respectively. Next, all of the exact matches were found and counted. For an exact match, the left and right boundaries had to match entirely.

Table 3. A summary of the inter-annotator agreement

Annotator	**# of Terms**	**Paired Annotators**	**# of Terms in Common**	***F* Score**
A	604	A-B	575	95.99%
B	594	A-C	576	96.16%
C	594	B-C	584	98.32%

Table 3 summarizes the inter-annotator agreement. Annotators A, B and C respectively identified 604, 594 and 594 terms independently. The total number of commonly identified terms is given for paired annotators. All the *F* scores for each paired annotators all above 95%, suggesting a high level of inter-annotator agreement. The results suggest that a high level of agreement is possible by training and by referring to the annotation guideline. Such a finding shows that trained annotators can achieve a high level of consistency even without expert domain knowledge, a finding that is contrary to the past experience that extensive training is needed for consistent annotation of terms in specialized domains such as biochemistry and medicine.

After the inter-annotator agreement test, the three annotators carried out the actual annotation and met to discuss the uncertain situations when necessary. Finally, the annotated corpus was manually validated by one annotator with the help of online resources and specialized dictionaries.

3 Syntactic Features of NP Constructions

In this section, we present some initial descriptive statistics and chart the distribution of NP constructions across different text categories and domains. We will first explain how we retrieve the syntactic functions of NPs according to the tree structure, followed by a description of the basic statistics of NP constructions in the corpus. We shall then present the preliminary observations of the syntactic features of NPs that are marked as terms.

3.1 A general description of NP constructions by category and domain

As explained in Section 2.1.2, every NP is assigned a function label and additional attributes if necessary. To count the frequency of NP constructions in trees is straightforward in most cases except for two scenarios, where the functions are labeled as `CJ` (conjoin; see Fig. 3) and `DEFUNC` (appositive NP that does not perform any syntactic function; see Fig. 4). In Fig. 3, the direct object NP is described by the attribute coordn, indicating the presence of a coordinated construction whose conjoins are marked as CJ. In such a scenario, a CJ will inherit the function of its mother node and be counted as a separate OD NP. Therefore, instead of counting one OD and two CJ functions, we count two OD functions for the NPs in Fig 3. Similarly, NPs with DEFUNC labels are also relocated and assigned the function label of the governing NP. See Fig. 4 for an example, where DEFUNC NP is treated as SU NP. In this particular case, instead of one DEFUNC and one SU, two SU functions are counted.

```
OD NP(coordn)
 CJ NP()
  DT DTP()
   DTCE ART(def) {the}
  NPHD N(com,plu) {gods}
 COOR CONJUNC(coord) {and}
 CJ NP()
  NPHD N(com,plu) {customs}
```

Fig. 3 – An example of CJ NP

```
SU NP()
 DT DTP()
  DTCE ART(def) {the}
 NPHD NADJ(sing) {unconscious}
DEFUNC NP(appos)
 NPHD PRON(ref,sing) {itself}
```

Fig. 4 – An example of DEFUNC NP

Table 4. Summary of NP constructions

	AHUM	**ASOC**	**ANAT**	**ATEC**	**PHUM**	**PSOC**	**PNAT**	**PTEC**
Function	**Freq**	**Freq**	**Freq**	**Freq**	**Freq**	**Freq**	**Freq**	**Freq**
A	26	40	14	66	54	61	60	40
AJPR	0	1	13	10	7	1	3	4
AVPR	3	6	9	3	16	11	11	8
CO	15	3	11	8	21	9	8	6
CS	215	147	144	184	260	180	190	172
DT	210	64	13	32	174	98	69	58
ELE	176	77	97	193	242	52	60	200
FOC	17	4	4	6	4	8	9	3
NPPO	250	150	419	237	246	59	32	99
NPPR	1	10	23	21	15	13	14	12
OD	850	806	634	778	924	951	812	947
OI	15	13	1	7	31	27	12	9
PC	3138	2982	3301	2834	3060	2585	2807	2356
PMOD	0	0	4	2	4	2	4	1
PROD	1	4	0	1	1	1	1	4
PRSU	33	53	39	37	29	55	32	42
SU	1685	1640	1626	1597	1986	1957	1859	1850
Total	**6635**	**6000**	**6352**	**6016**	**7074**	**6070**	**5983**	**5811**

With this treatment of conjoin and appositive NPs, NP constructions in all the eight subject domains were retrieved and summarized in Table 4. As can be observed in Table 4, there is an uneven distribution of 17 different functions of NPs across domains. In general, NPs seem to occur most frequently at the position of PC in all the domains, followed by SU and OD. Nevertheless, when we examine the functions by category and domain, we notice more interesting patterns. First, NPs in domains of academic writing tend to occur less frequently at the position of SU than those in their counterparts of popular writing. Second, domains in academic writing are more likely to have a comparatively higher occurrence of PC as a syntactic function than their counterparts in popular writing. They also tend to have fewer occurrences of OD.

3.2 A statistical description of term-NP constructions

When examining the distribution of term-NPs, we also related the CJ and DEFUNC functions to their mother nodes. Accordingly, the actual distribution of term-NPs across difference categories and domains were calculated and presented in Table 5.

Table 5. Summary of term-NP constructions

	AHUM	**ASOC**	**ANAT**	**ATEC**	**PHUM**	**PSOC**	**PNAT**	**PTEC**
Function	**Freq**	**Freq**	**Freq**	**Freq**	**Freq**	**Freq**	**Freq**	**Freq**
A	4	3	1	16	2	1	1	5
AJPR	0	0	2	4	0	1	0	1
AVPR	0	0	0	1	0	1	1	0
CO	12	1	10	5	7	4	4	4
CS	106	48	63	76	85	55	68	36
DT	140	29	8	20	73	54	40	35
ELE	16	35	40	47	56	14	42	92
FOC	12	1	3	4	2	0	6	2
NPPO	10	14	7	5	10	1	1	3
NPPR	0	7	14	9	2	5	11	6
OD	456	341	316	408	316	331	379	480
OI	5	5	0	45	6	4	4	2
PC	1637	1435	1886	1496	1043	982	1199	1082
SU	510	536	753	654	422	442	673	621
Total	**2908**	**2455**	**3103**	**2790**	**2024**	**1895**	**2429**	**2369**

Interesting features emerge from the initial frequency count. First, academic writing tends to have more terms than popular writing in both parameters (i.e. category and domain). In a broad sense, the total number of terms in academic writing is higher than that of popular writing. From the perspective of subject domains, individual domains belonging to academic writing tend to have more terms than their counterparts in popular writing. Such a result suggests that formal writing tends to contain more term candidates than informal writing. Second, science domains (i.e. NAT and TEC) tend to contain more terms than arts domains (i.e. HUM and SOC). It can be also noticed that the number of terms in AHUM is higher than that of ATEC, and it is understandable since AHUM has the highest number of NPs among the domains in academic writing. Third, across the eight domains term-NPs seem to appear most frequently at the position of PC, followed by SU and OD. Fourth, it would be easy to make a contrastive study on certain syntactic functions across the eight domains. For example, terms are more likely to occur at the position of A in ATEC when compared with the other seven domains, and they are more likely to appear at the position CS in AHUM when examined across domains. Such information can be taken as a flexible value in assigning weights to syntactic functions in accordance with particular domains in ATR.

It is worth mentioning that syntactic labels at the phrasal level can be further classified at the clausal level. For example, a considerable number of NPs occur at the position of PC, which should be related to its mother node, namely PP, whose functions could be analyzed differently as A PP and NPPO PP, revealing further variations of use across the eight categories.

4 Conclusion

In this paper, we presented a new corpus resource that has been constructed specially for the study of the syntactic characteristics of terminologies for a linguistically motivated description of terms and their internal structures. The corpus is based on the British component of the International Corpus of English, comprising four parallel subject domains from two text categories (i.e. academic vs. popular prose) with a total of about 200,000 running word tokens. It is richly annotated at lexical, grammatical, syntactic, and terminological levels. It is parameterized according to both text categories and subject domains. We first described the construction of the corpus, including the selection of the corpus material, the annotation schemes for grammar and syntax, and an inter-annotator analysis of the annotation of terms. We then described the corpus resource by reporting some of our initial empirical observations of NP constructions and term-NP constructions. Interesting patterns were observed in terms of syntactic distribution of NPs and term-NPs across different categories and domains. In particular, term-NPs show observable difference across different categories and domains. In other words, the corpus resource can provide an analytical framework for the study of relations between terminological use and text types as well as subject domains.

Acknowlegement

The work described in this paper was supported partially by research grants (Nos 7002190, 7200120 and 7002387) from City University of Hong Kong. The authors would like to thank the reviewers for their valuable comments and suggestions.

References

ANANIADOU, S., ALBERT S. & SCHUHMANN D. (2000). Evaluation of Automatic Term Recognition of Nuclear Receptors from MEDLINE. *Genome Informatics* 11. p. 450-451.

AUBIN, S. & HAMON T. (2006). Improving Term Extraction with Terminological Resources. In *FinTAL 2006, LNAI 4139*. p. 380-387.

CORBETT, P., BATCHELOR, C. & TEUFEL, S. (2007). Annotation of Chemical Named Entities. In *Proceedings of BioNLP 2007: Biological, translational, and clinical language processing*. p. 57-64.

DEMETRIOU, G. & GAIZAUSKAS, R. (2003). Corpus resources for development and evaluation of a biological text mining system. In *Proceedings of the Intelligent Systems for Molecular Biology (ISMB) Workshop on Text Mining (BioLINK2003)*.

EUMERIDOU, E., NKWENTI-AZEH, B. & MCNAUGHT, J. (2004). An Analysis of Verb Subcategorization Frames in Three Special Language Corpora with View towards Automatic Term Recognition. *Computers and the Humanities,* 38. p. 37-60.

FANG, A.C. (1996). AUTASYS: Automatic Tagging and Cross-Tagset Mapping. In S. GREENBAUM Ed. p. 110-124.

FANG, A.C. (2000). From Cases to Rules and Vice Versa: Robust Practical Parsing with Analogy. In *Proceedings of the Sixth International Workshop on Parsing Technologies*. p. 77-88.

FANG, A.C. (2006). Evaluating the Performance of the Survey Parser with the NIST Scheme. In *LNCS 3878: Computational Linguistics and Intelligent Text Processing,* ed. by A. Gelbukh. Berlin Heidelberg: Springer-Verlag. P. 168-179.

FANG, A.C. (2007). *English Corpora and Automated Grammatical Analysis*. Beijing: The Commercial Pess.

FRANTZI, K., ANANIADOU, S. & MIMA, H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal of Digital Libraries,* 3(2). p.117-132.

GREENBAUM, S. (1996). *Comparing English World: The International Corpus of English*. Oxford: Oxford University Press.

HA, L. A., MITKOV, R. & CORPAS, G. (2008). Mutual Terminology Extraction Using a Statistical Framework. In *Procesamiento del lenguaje natural*. N. 41. p. 107-112.

KIT, C. & LIU, X. (2008). Measuring Mono-word Termhood by Rank Difference via Corpus Comparison. *Terminology* 14:2. p. 204-229.

KOLÁŘIK, C. & KLINGER, R. (2008). Chemical Names: Terminological Resources and Corpora Annotation. *LREC 2008 Workshop: Building and Evaluating Resources for Biomedical Text.* p. 51-58.

L'HOMME, M.C. (2002). What can Verbs and Adjectives Tell us about Terms? In *Proceedings of Terminology and Knowledge Engineering (TKE 2002)*. p. 65-70. France.

MORGAN, A.-A., HIRSCHMAN, L., COLOSIMO, M., YEH, A.-S. & COLOMBE, J.-B. (2004). Gene Name Identification and Normalization using a Model Organism Database. *Journal of Biomedical Informatics*, 37(6). p. 398-410.

NAKAGAWA, H. & MORI, T. (2003). Automatic term recognition based on statistics of compound nouns and their components. *Terminology,* 9 (2). p. 201-219.

NENADIC, G., SPASIC, I., & ANANIADOU, S. (2005). Mining Biomedical Abstracts: What's in a Term? In K.-Y. Su, J. Tsujii, J.-H. Lee and O. Kwong (Eds.) *Natural Language Processing – IJCNLP 2004 First International Joint Conference, Lecture Notes in Computer Science vol. 3248*. p. 797-806.

OHTA, T., TATEISI, Y., MIMA, H. & TSUJII, J. (2002). GENIA Corpus: An Annotated Research Abstract Corpus in Molecular Biology Domain. In *Proceedings of the Human Language Technology Conference*, San Diego, USA.

PAZIENZA, M.T., PENNACCHIOTTI, M. & ZANZOTTO, F.M. (2005). Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. *StudFuzz* 185. p. 255-279.

RODRÍGUEZ, B., MARIO, F., NOYA, E. D., OTERO, P.G., MARTÍNEZ, M.L., MATO, E. M.M., ROJO, G. SANTALLA DEL RÍO, M.P., & DOCÍO, S.S. (2007). A Corpus and Lexical Resources for Multi-word Terminology Extraction in the Field of Economy in a Minority Language. In *Proceedings of the 3rd Language and Technology Conference*. p. 359-363.

VILLE-OMETZ, F., ROYAUTÉ, J. & ZASADZINSKI, A. (2007). Enhancing automatic recognition and extraction of term variants with linguistic features. *Terminology* 13 (1). p. 35-59.

VLACHOS, A. & GASPERIN, C. (2006). Bootstrapping and Evaluating Named Entity Recognition in the Biomedical Domain. In *Proceedings of BioNLP in HLT-NAACL*. p. 138-145.

Appendix: A Complete List of Parsing Symbols

`A`	Adverbial
`ADJ`	Adjective
`ADV`	Adverb
`AJHD`	Adjective phrase head
`AJP`	Adjectve phrase
`AJPO`	Adjective phrase postmodifier
`AJPR`	Adjective phrase premodifier
`ANTIT`	Anticipatory *it*
`ART`	Article
`AUX`	Auxiliary
`AVB`	Auxiliary verb
`AVHD`	Adverb phrase head
`AVP`	Adverb phrase
`AVPO`	Adverb phrase postmodifier
`AVPR`	Adverb phrase premodifier
`CF`	Focus complement
`CJ`	Conjoin
`CL`	Clause
`CLEFTIT`	Cleft *it*
`CLOP`	Cleft operator
`CO`	Object complement
`CONJUNC`	Conjunctor
`CONNEC`	Connector
`COOR`	Coordinator
`CS`	Subject complement
`CT`	Transitive complement
`DEFUNC`	Detached function
`DISMK`	Discourse marker
`DISP`	Disparate coordination
`DT`	Determiner
`DTCE`	Central determiner
`DTDE`	Deterred determiner
`DTP`	Determiner phrase
`DTPE`	Pre-determiner
`DTPO`	Determiner postmodifier
`DTPR`	Determiner premodifier
`DTPS`	Post-determiner
`ELE`	Clause element
`EXOP`	Existential operator *there*
`EXTHERE`	Existential *there*
`FOC`	Focus
`FRM`	Formulaic expression
`GENF`	Genitive function
`GENM`	Genitive marker
`IMPOP`	Imperative operator
`INDET`	Indetermined
`INTOP`	Interrogative operator
`INVOP`	Inversion operator
`LIM`	Limitor
`LK`	Linker
`MVB`	Main verb
`N`	Noun
`NADJ`	Nominal adjective
`NONCL`	Non-clause
`NOOD`	Notional object
`NOSU`	Notional subject
`NP`	Noun phrase
`NPHD`	Noun phrase head
`NPPO`	Noun phrase postmodifier
`NPPR`	Noun phrase premodifier
`NUM`	Numeral
`OD`	Direct object
`OI`	Indirect object
`OP`	Operator
`P`	Prepositional
`PARA`	Paratactic
`PC`	Prepositional complement
`PMOD`	Preposition premodifier
`PP`	Prepositiional phrase
`PRED`	Predicate
`PREDG`	Predicate group
`PREP`	Preposition
`PROD`	Provisional object
`PROFM`	Pro-nominal form
`PRON`	Pronoun
`PRSU`	Provisional subject
`PRTCL`	Particle
`PS`	Stranded preposition
`PU`	Parsing unit
`PUNC`	Punctuation
`REACT`	Reactional signal
`SBMO`	Subordinator phrase premodifier
`SU`	Subject
`SUB`	Subordinator
`SUBHD`	Subordinator phrase head
`SUBP`	Subordinator phrase
`TO`	Infinitive *to*
`V`	Verb
`VB`	Verbal
`VP`	Verb phrase